Published in final edited form as:

Biomark Med. 2009 October ; 3(5): 577–588. doi:10.2217/bmm.09.46.

Identification of a β -casein-like peptide in breast nipple aspirate fluid that is associated with breast cancer

Edward R Sauter^{1,†}, Wade Davis^{2,3}, Wenyi Qin¹, Sarah Scanlon⁴, Brian Mooney^{4,5}, Karen Bromert⁴, and William R Folk⁴

¹Department of Surgery, University of North Dakota School of Medicine & Health, Sciences, 501 N. Columbia, Road, Stop 9037, Grand Forks, ND 58201, USA

²Department of Health, Management & Informatics, University of Missouri, USA

³Department of Statistics, University of Missouri, USA

⁴Department of Biochemistry, University of Missouri, USA

⁵The Charles W Gehrke, Proteomics Center, University, of Missouri, USA

Abstract

Aims—Nipple aspirate fluid was collected prospectively from women scheduled for diagnostic breast surgery in order to determine protein masses associated with breast cancer, subsets of women with a unique proteomic profile and a breast cancer predictive model.

Materials & methods—Breast nipple aspirate fluid was collected preoperatively in 163 breasts from 125 women and analyzed for changes in cell morphology and by SELDI-TOF mass spectrometry over approximately a 44 kDa range (1.5–45 kDa) using IMAC30, CM10 and Q10 ProteinChips.

Results—Considering all samples, 16 protein masses were associated with the presence of cancer, the most discriminating being 3592, 6570/6580 and 15870 Da. Excluding women with pathologic nipple discharge or those with a papilloma identified an additional protein of 6383 Da. The best cancer detection models included Breast Imaging Reporting and Data System, age, and either the 4262 (best sensitivity: >87%) or 3592 (best specificity: >94%) peak. MALDI-TOF mass spectrometry demonstrated the 3592 peak, which was most discriminating in many of our cancer prediction models, to be a β -casein-like peptide.

Conclusion—Differential nipple aspirate fluid proteomic expression exists between women with/without breast cancer. The most discriminating protein identified is a β -casein-like peptide not previously described. Combining proteomic and clinical information, which are available before surgery, optimizes the prediction of which women have breast cancer.

Keywords

biomarker; breast cancer; early detection; nipple aspirate fluid; proteomics

© 2009 Future Medicine Ltd

[†]Author for correspondence: Department of Surgery, University of North Dakota, School of Medicine & Health, Sciences, 501 N. Columbia Road, Stop 9037, Grand Forks, ND 58201, USA, Tel.: +1 701 777 2284; Fax: +1 701 777 4158; esauter@medicine.nodak.edu.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Suspicious lesions identified using imaging (mammography, breast ultrasound or breast MRI) require tissue collection and histopathologic assessment through diagnostic needle, core or surgical biopsy, which are painful, and only approximately 15–25% of the procedures demonstrate malignancy [1,2]. SELDI-TOF (Ciphergen Biosystems, CA, USA) mass spectrometry (MS) has been investigated for the diagnosis, prognosis and therapeutic monitoring of cancer. The technology utilizes a combination of MS and chips with affinity to specific protein types to facilitate protein profiling of complex biological mixtures. It has higher throughput capability than 2D polyacrylamide gel electrophoresis [3]. Preliminary studies using SELDI-TOF MS assessment of body fluids to detect breast cancer have focused primarily on serum and plasma. A concern with the analysis of serum or plasma is the contribution of every bodily organ to the circulating protein pool, leading to the dilution of potential cancer predictive protein/peptide markers. We [4,5] and others [6–9] have attempted to identify breast-specific cancer predictive proteins in nipple aspirate fluid (NAF) in studies generally of limited size and with different participant enrollment and peak discrimination criteria.

In two pilot studies we performed SELDI-TOF MS on noninvasively collected NAF samples and identified five protein ion masses in the first (6500, 8000, 15,940, 28,100 and 31,770 Da) and four in the second (5200, 11,880, 13,880 and 33,400 Da) that were associated with breast cancer [4,5]. The 8000, 15,940 and 31,770 Da proteins are mono-, di- and tetra-meric forms of hemoglobin [4], and therefore removed from consideration as biomarkers. Although promising, both studies were limited by the method of peak determination (visual inspection of the spectra or subtraction, i.e., comparison of peak area) of mass spectra using Ciphergen ProteinChip Software. Peak determination in our current report used quantitative, objective and reproducible statistical algorithms. Additionally, a second generation of ProteinChips (cationic [CM10], anionic [Q10] and immobilized metal affinity chromatography [IMAC30]) with the potential for greater protein peak discrimination were used in the current report.

We conducted a biomarker study in women requiring diagnostic breast surgery. Cytologic analysis was performed. Clinical variables, including Breast Imaging Reporting and Data System (BI-RADS) information, were also included in our breast cancer predictive models [10].

Patients & methods

Overview

Women aged 18 years and over requiring diagnostic breast surgery were recruited from breast evaluation clinics in the University of Missouri Health System and prospectively enrolled in an Institutional Review Board approved protocol. Most but not all subjects underwent mammography. For those who underwent mammography, BI-RADS categories ranged from 0 to 5, including categories 4 (suspicious) and 5 (malignancy likely) [11]. Subjects may have undergone needle but not surgical biopsy prior to NAF collection. Subjects could not have been receiving chemotherapy or radiation therapy at the time of nipple aspiration. Pathologic (single duct or grossly bloody discharge) nipple discharge (PND) requires diagnostic breast surgery to exclude malignancy. We previously observed [12] that biomarker expression may be different in breasts with PND requiring surgery compared with those without. For this reason, we determined if SELDI-detected biomarkers would differ if we compared the entire group (samples from breasts with and without PND) to only samples from breasts without PND.

Specimen procurement, preparation & interpretation Aspiration technique

After nipple cleansing with alcohol, fluid from the breast that subsequently underwent surgery was aspirated by a trained physician or nurse clinician using a modified breast pump [13] and collected in capillary tubes. Half of the fluid (generally 1–5 μ l per tube) was collected in 50 μ l capillary tubes and placed on ice until being snap frozen at -80°C . The remainder was placed into Shandon fixative (ThermoShandon, PA, USA) for cytologic analysis.

Histologic & cytologic evaluations

The definitive diagnosis of breast cancer was made from histopathologic review of the diagnostic biopsy tissue specimen by the pathologist assigned to the case. Cytology evaluation of NAF was performed by a cytopathologist with experience in evaluating cytology. The cytopathologist was blinded to the tissue diagnosis. Each fixed sample was centrifuged onto six glass slides for cytologic analysis. Papanicolaou-stained smears were examined in blinded fashion and classified as inadequate, normal, ductal hyperplasia, mild-to-moderate atypia, marked atypia or malignant cells present [14,15].

Protein profiling

Specimen preparation & analysis—Blinded sample analysis was performed in duplicate using SELDI-TOF. NAF was diluted in 100 μ l of $1\times$ phosphate-buffered saline, 0.1% Triton-X 100. Protein concentrations were determined using the Pierce bicinchoninic acid kit (Pierce Chemical Co., IL, USA) and then further diluted to a final concentration of 3.2 mg/ml. A cocktail of protease inhibitors (aprotinin, pepstatin A, leupeptin and 4-[2-aminoethyl] benzenesulfonyl fluoride hydrochloride) was added to each 40 μ l NAF sample, mixed with 60 μ l urea buffer (9 M urea, 2% CHAPS, 50 mM Tris pH 9), and applied to IMAC30, CM10 and Q10 ProteinChips. Pre-equilibration and washing steps were carried out as instructed by the manufacturer.

Assessment of consistency of the SELDI-TOF technique—External standards were added to all NAF samples: bovine ubiquitin (8564 Da); bovine insulin (5733 Da); human b-endorphin (3465 Da); human ACTH (2933 Da); porcine dynorphin A (2147 Da); fibrinopeptide (1570 Da); human angiotensin (1296 Da), bovine cytochrome C (12,230 Da); bovine superoxide dismutase (15,591 da); equine myoglobin (16,950); bovine lactoglobulin (18363 Da); and horseradish peroxidase (43,240 Da). Calibration checks were performed with each sample run. Mass calibration was always maintained to ± 500 ppm. Each spectrum was evaluated on a set of three statistics (quality, peak and retain) to determine whether it was of sufficient overall quality [16,17]. If the spectrum did not meet minimum standards, it was not analyzed. We also computed correlations between replicate spectra as another measure of reliability. The best spectrum was selected for each sample [17] based strictly on the quality of the spectrum, and not on any clinical covariates or disease status.

Identification of promising protein peaks using SELDI-TOF—Spectrum peaks were detected using a published algorithm [16]. Spectra of the same chip type and laser intensity were processed together to identify common peaks as potential biomarkers. Potential biomarkers were computed as previously described [18]. Statistical tests were used to determine which peaks were significantly different between cancerous and noncancerous breasts, based both on peak intensity and the presence/absence of a given peak.

Protein identification of protein mass peaks that differentiate cancer from benign breasts

Purification of the peaks from NAF—Nipple aspirate fluid diluted to 3.2 mg/ml was processed using ProteinChip IMAC spin columns (Bio-Rad, Hercules, CA, USA). NAF was

loaded on the prepared column, unbound proteins removed by centrifugation, the column washed and specifically bound proteins/peptides eluted with 10% formic acid. Six elution fractions were collected from each column.

MALDI mass spectrometry—An aliquot of each elution was mixed with an equal amount of α -cyano hydroxycinnamic acid matrix, spotted onto a MALDI target and peptides and matrix allowed to co-crystallize. The instrument (4700 Proteomics Analyzer, Applied Biosystems, CA, USA) was operated in positive ion mode and spectra were acquired over a mass range of 2000 to 4000 Da. Six peptides of known mass (Applied Biosystems) were used to calibrate the instrument in MS mode. Calibration was also conducted using fragment ions of 1570.7 Glu-1-Fibrinopeptide B on all six calibrant spots. Following an MS scan of each sample, the SELDI-identified ions were selected for MS/MS acquisition. An NCBI database search was conducted limited to mammals (last updated 23 September, 2007) using the 'MS/MS search' function of GPS Explorer (v 3.6) software, which is integrated with the MASCOT V2.1 [101] search algorithm. Search parameters allowed: nitric oxide enzyme and methionine oxidation as a variable (possible) modification. MS mass tolerance was 150 ppm and MS/MS mass tolerance was 0.2 Da.

Electrospray mass spectrometry—Instrument calibration (Qstar pulsar i) was achieved as instructed (Applied Biosystems). The purified NAF sample (as described earlier) was loaded into a nanospray emitter (Proxeon Biosystems, CA, USA) and introduced into the instrument using an ionspray voltage of 900 V. MS spectra were acquired in positive ion mode across 400–2000 m/z (mass/charge). Peptide ions of interest were selected for MS/MS acquisition and data acquired in MCA mode with a collision energy of 50. Database searches were conducted using BioAnalyst and MASCOT. *De novo* sequencing – manual interpretation of the fragmentation (MS/MS) spectra – was conducted on the quintuple-charged ion of the 3603 Da peptide. Database matches to the *de novo* sequence were identified using the basic local alignment search tool for proteins (BLASTP) utility of the National Center for Biotechnology Information (NCBI).

Statistical methods

SELDI spectra were read into the software package R (Version 2.2.1, R Foundation for Statistical Computing, Vienna, Austria). Low-level preprocessing (baseline removal and normalization, based upon total ion current [TIC] for a given chip) was performed using the PROcess package. Locally weighted regression was used to estimate spectrum baseline. All spectra of a given chip type/laser intensity used the same parameter values in the baseline removal algorithm. TIC was computed for each spectrum independently, then every spectra of similar chip type/laser intensity was multiplied by the appropriate constant so that the TIC of every spectra was equal to the overall median for that chip type/laser intensity. TIC was computed using only masses above the mass deflector setting. Normalization was carried out separately for each chip type/laser intensity.

Two types of hypothesis tests (normalized spectral intensities–Wilcoxon–Mann–Whitney and peak presence or absence–Fisher's exact test) were conducted on identified candidate biomarkers. Although hypothesis testing was not our primary aim, but rather prediction, we computed both raw p-values as well as adjusted p-values based on the Benjamini–Hochberg correction. This approach was pursued since intensity values may not be optimal for biomarker determination owing to experimental noise [19]. The area under a peak, or presence/absence may be more reflective of the amount of protein present than peak intensity, especially at higher masses [20]. Groups were compared based only on the presence or absence of a peak at a given mass.

Some covariates are at the subject level (e.g., age and family history) while many others are at the breast level (e.g., pathologic and cytologic diagnosis of cancer or benign, proteins from NAF and PND). Importantly, most women with cancer in one breast do not develop contralateral breast cancer during their lifetime. As such, we chose to consider covariates at the breast level as independent observations. We acknowledge that this perspective is a potential limitation of the study.

The entire set of identified peaks was used as inputs into several different classification algorithms, where the predictor variables were peak intensity. Each of the classification algorithms include a variable selection stage in which the set of potential biomarkers is selected from the set of all peaks found across all chips and laser intensity combinations. Multivariate binary (presence/absence) predictive models were not considered owing to high-dimensionality relative to the number of subjects in the study. Only cases with complete spectral data (e.g., data from IMAC30, CM10 and Q10 ProteinChips) were used in build-in predictive models.

Results

Subject population

A total of 132 women were enrolled. NAF samples were successfully collected from 125 (95%) enrolled women who ranged in age from 22 to 95 years (median: 53 years) (Table 1). NAF samples were evaluated from one breast in 87 subjects, and from both breasts in 38 subjects, thus a total of 163 breasts were evaluated. BI-RADS category was recorded for each breast (Table 2). In total, 63% of women were postmenopausal and 24% had a first-degree relative with breast cancer. One woman had bilateral breast cancer. A total of 37 women presented with PND with seven of these having unilateral breast cancer. A total of 20 women presented with a papilloma, the most common cause of PND, and none of these women had breast cancer. The number of breasts with each pathologic diagnosis were as follows: 111 benign (including 20 papillomas), four atypical duct hyperplasia, 11 ductal carcinoma *in situ* (DCIS) and 37 invasive breast cancer (IBC).

We compared clinical factors (age, menopausal status, history of prior breast biopsies, first degree relatives with breast cancer, current birth control pill, hormone-replacement therapy use and parity) and breast cancer status. The percentage of women who were postmenopausal, participant age and number of first-degree relatives with breast cancer were greater in women with cancer than in those without (Table 1). There were no significant differences in the other clinical factors based on breast cancer status. Two subjects underwent needle biopsy prior to NAF collection – the remaining subjects did not. Needle biopsy did not significantly alter biomarker results. Computed p-values in Table 1 used Fisher's exact test except for age, which was analyzed using a two-sample t-test.

Results of NAF cytology

In total, 37 samples showed IBC and 11 indicated DCIS; two of the IBC and two of the DCIS samples demonstrated cytology that was either frankly malignant or highly suspicious for cancer. Atypia was found in an additional three IBC and four DCIS samples. Thus, only five out of 37 (14%) IBC samples compared with six out of 11 (55%) DCIS samples demonstrated cytologic atypia, cells suspicious for cancer or frankly malignant cells.

One of 162 NAF samples that underwent cytologic review (cytology was unavailable for one NAF sample) and was diagnosed as malignant came from a breast in which the removed tissue did not demonstrate evidence of DCIS or IBC. Thus, the false-positive rate of NAF cytology was less than 1%.

Protein biomarkers in women scheduled for diagnostic biopsy that are associated with breast cancer—Spectral data were collected from 163 unique breasts using IMAC chips, 157 using Q10 chips and 141 using CM10 chips. At least one spectra for every breast sample was usable for IMAC and CM10, while all but three Q10 spectra passed quality screening. The total number of spectra differ slightly between chip types owing to a staggered roll-out period between chip types. We also computed the correlation coefficient between each pair of duplicate spectra, which resulted in 922 correlation coefficients (461 duplicate pairs each shot at two different laser intensities). The median correlation coefficient between duplicates was 92% (interquartile range: 9%). Four samples were run in duplicate 5–10 months after they were initially analyzed to see if the strong duplicate correlation was still observable. The average difference in the correlations between the two time points was miniscule (−0.023), further confirming the reproducibility of the spectra. When we divided samples by whether or not they had cancer (defined as DCIS or IBC), five proteins (3592, 15334, 15135, 15157 and 15870 Da) were associated with the presence of breast cancer and 11 (6570, 6580, 6710, 11747, 13005, 13207, 14698, 15427, 15447, 22434 and 22512 Da) with the absence of cancer in univariate analysis obtained from one or more chips (Table 3). Considering protein masses within 1% of each other to be possibly related, the 6570/6580, 6710 and 15870 Da peaks were possibly related to proteins we previously identified as associated with breast cancer [4,5], whereas peaks not found in our earlier studies at 11747, 13005, 13207, 14698, 15135/15157, 15334, 15427/15447, 22434 and 22512 Da were also identified.

The best sensitivity (~67%) mass (15870) was approximately threefold more likely (odds ratio: 3.33) to be present in cancerous than non-cancerous breasts. The most specific (84%) mass was 3592. Mass 6570 was fivefold more likely (odds ratio: 0.19) to be present in noncancerous than cancerous breasts.

NAF protein profile based on presentation without PND—After excluding women with PND, seven additional peaks (6383, 6441, 6621, 11719/11739 and 11916) discriminating cancerous from non-cancerous breasts were detected (Table 4 & Figure 1) and four peaks (11747, 15334, 15447 and 22434) that differentiated cancer from noncancer when all samples were considered were absent when PND samples are excluded. In the no PND group, 6383 had the best sensitivity (88%) and 3592 (87%) the best specificity.

Identification of a β -casein-like peptide—The SELDI software averaged 3592 Da cancer-specific peak (Figure 2A) was corroborated by MALDI TOF-TOF MS analysis to be two peaks (3585.5 and 3603.6 Da; Figure 2B). In some NAF samples both peaks were present, whereas in others only the 3603.6 peak was observed. *De novo* sequencing of the 3602 ion yielded four sequences that were searched against the NCBI database limited to humans. Electrospray ionization-based QTOF analysis (Applied Biosystems QSTAR Pulsar I) yielded two major ions (Figure 3A), one of which yielded by *de novo* sequencing (Figure 3B) a putative sequence of KKVEKVKHEDQQQGEXEHQDKIYXXXXPQP, with the best match upon searching NCBI database limited to humans yielded β -casein (GI number 288098) with an excellent E value (4e-18). At position 33–62 within the β -casein sequence, the following peptide KKVEKVKHEDQQQGEDEHQDKIYPSFQPQP matches the molecular mass of the 3603 Da ion perfectly (with loss of ammonia from the true sequence molecular weight). There are similarities to the sequence derived from the MALDI TOF-TOF data but not complete agreement. The sequence derived therefore is the best possible match based on the actual MS/MS data and a database match. The ions present in the MS/MS spectrum match this sequence very well. Consistent with the previous SELDI and TOF-TOF data, the QSTAR MS/MS spectrum shows loss of water and ammonia (−18 and −17 Da, respectively), further supporting the assertion that the 3602 and 3585 Da MALDI TOF-TOF demonstrated peaks are related by loss of ammonia.

Multivariate analyses

Our list of potential predictors consisted of the intensity of SELDI-detected peaks from all three chips plus clinical variables typically available prior to diagnostic breast surgery, including age, race, cytology, BI-RADS category, parity, menopausal status, birth control use, PND and hormone-replacement therapy use. A variety of predictive modeling techniques (step-wise logistic regression, stepwise linear and quadratic discrimination, and classification and regression trees) were used. These methods incorporate a built-in variable selection step, which is important because the cancer and no cancer groups differ with respect to some clinical variables (Table 1), such as age. Since these clinical variables were included in the list of possible predictor variables, these models are implicitly adjusted for differences between the groups based on the clinical variables selected by a given model. For the logistic regression models, the stepwise procedure was based on the commonly used Akaike information criterion. Since Akaike information criterion is not possible for the other modeling approaches used (e.g., linear discrimination), variables had to improve the classification accuracy by at least 1% to be included in the model.

Table 5 summarizes the classification performance of the various algorithms in terms of sensitivity, specificity and overall classification accuracy using the leave-one-out cross-validated (LOOCV) technique. In LOOCV, an observation is 'left out' from the dataset, the classification model is trained using the remaining data and then the model is used to predict the left out observation. This is repeated for every data point, and the resulting statistics are calculated based on the performance on the left out data. LOOCV has been shown to be nearly unbiased method for estimating the classification error rate (and therefore accuracy) and is appropriate for moderate sample sizes such as ours [21]. The LOOCV was used at each step in the stepwise procedure discussed previously.

Table 5 also includes the variables were contained in the best model for each method using only clinical data, SELDI data and SELDI in combination with clinical data. In a minority of cases, protein masses (e.g., 4129 and 4262) may be present in the multivariate modeling illustrated in Table 5 but not in Table 3 & Table 4, which are based on univariate analyses. In multivariate modeling, the predictor variables considered consisted of all SELDI-detected peaks plus clinical variables. Thus, peak 4262 was not significant by itself in univariate analysis, but contains predictive information when combined with BI-RADS category and age.

The best predictive models consisted of the clinical variables, BI-RADS and age, along with the SELDI peaks at 4262 or 3592 Da. The cross-validated sensitivities of the better models were greater than 80% and the specificities greater than 93% (Table 5).

Discussion

Whenever a new technology is introduced, it is essential to evaluate its reliability. We were encouraged to observe a high level of correlation between replicates (median correlation: 92%, interquartile range: 9%), and that we were able to include virtually all samples for analysis. As observed in the past [15], we found NAF cytology to be highly specific but not very sensitive in determining if a breast contains cancer. In the end, cytology did not make it into any of the models for cancer detection.

Compared with the entire cohort, after removing samples from women with PND, the sensitivity of most peaks increased while the specificity decreased, likely because many more normal breasts were excluded with PND than cancerous breasts with PND. For the non-PND group protein mass 6383 had the best sensitivity, while 3592 had the best specificity for both the entire and non- PND groups. For both groups the mass whose

absence was associated with the highest odds of cancer was 6570, whereas the mass whose presence was associated with the highest odds of cancer was 15870. Based on an earlier study, it is likely that the 15870 peak represents hemoglobin- β [4]. As can be seen for the 15870 peak in Table 3 & Table 4, the association of this peak with breast cancer decreased when PND samples were eliminated. Of the eight models for which SELDI proteins were considered (Table 5), 3592 was independently associated with cancer in five models.

We compiled a list of previous studies we are aware of that performed SELDI-TOF analysis of NAF obtained from women with and without breast cancer (Table 6). Sample size was under 100 in all studies, most cases compared healthy women (rather than those requiring surgery) with women with breast cancer, and the timing of NAF collection relative to the surgical procedure varied. Most studies were conducted retrospectively, which increases the chance for sample heterogeneity and bias. Some studies removed abundant proteins prior to SELDI analysis, different ProteinChip(s) were used, different NAF total protein concentrations were chosen for SELDI analysis, the criteria used to classify a peak as associated with cancer or no cancer and the approach to statistical analysis varied among the studies. All of these differences likely influenced the proteins that were determined to be associated with cancer or no cancer. The current study included a homogeneous population of women who all had an abnormality requiring diagnostic breast surgery. All samples were collected prior to excisional biopsy or mastectomy. Validation of our findings using a similar cohort of women, in other words, those requiring diagnostic breast surgery, will be essential prior to clinical use, to confirm that the findings are reproducible.

Since subjects were on average younger in the no-cancer than the cancer group, we constructed an age*peak interaction term in each of our multivariate statistical models. If the term entered a model, it would imply that the impact of a certain peak (e.g., 3592) on the odds of cancer differs depending on age. In no model did the age*peak interaction term enter a model. To further investigate this, for each final model selected by the stepwise procedure, we forced the corresponding age*peak term into the model. Model classification performance (% correct) decreased in most cases, remained the same for a few others and improved in no cases, therefore the results shown in Table 5 do not include interaction terms.

Some but not all of the peaks that we identified in our previous two studies [4,5] were identified in the current analysis as being associated with breast cancer. Peaks of similar size include 6570/6580 Da (compared with 6500 Da) and 15870 Da (compared with 15940 Da); 6710 Da is close in size to 6650 Da, which we previously identified, but 6650 Da was not significantly associated with breast cancer [4] in the earlier study. We did not identify 3592 Da in our earlier studies, but herein note that it consists of two (3585 and 3602) ions, and describe its identification as a peptide fragment of β -casein. Peptide fragmentation (tandem MS-MS/MS) patterns for the two ions were quite similar suggesting that these peptide sequences are related, possibly by loss of ammonia (-17 Da). This loss could result *in vivo*, during sample preparation, or even during MALDI TOF-TOF (or SELDI) acquisition (ionization). The fragment of β -casein identified in this study terminates in a proline residue. It is interesting that a so-called 'casomorphinase' has been shown to have this proline-specific cleavage (US patents 6251391 and 6808708) and casomorphins are fragments of the β -casein sequence 60–70 and bind to the μ -opoid receptor [22], which have opioid activity and can influence breast cancer proliferation [23]. Proline-specific proteases also play a role in other biological processes [24]. There was no correlation between number of live births, no and one or more live births, and the presence of the 3592 peak.

Finally, we wished to determine if protein profiling using SELDI-TOF analysis plus clinical variables would be associated with breast cancer. The optimal model was 92% specific and

87% sensitive in determining if a breast requiring surgery had cancer. The model was correct 91% of the time. By comparison, MRI was recently reported to be 77–100% sensitive and 81–97% specific in detecting breast cancer in high-risk women, with lower accuracy in detecting DCIS [25]. The SELDI (mostly nonhigh-risk women) and the MRI (high-risk women) populations differed, so comparisons of accuracy may not be appropriate. Nonetheless, our findings suggest that SELDI-TOF analysis of NAF plus clinical information known prior to surgery is quite specific in determining if a breast contains cancer. NAF analysis by SELDI [4] appears to be the best in the detection of DCIS, whereas MRI is better in the detection of more advanced disease. Proteins identified in NAF may be more predictive of DCIS than IBC because once invasive, the tumor obstructs the duct and grows away from the duct from which one is collecting the sample. These modalities, therefore, may be complementary in breast cancer screening, since they are best at detecting different stages of breast cancer. Confirmation of these findings will be pursued in forthcoming studies.

Conclusion

We prospectively collected NAF in women scheduled for diagnostic breast surgery. Protein expression differed among women with and those without PND. The proteins that best discriminated if a breast contained cancer were 3592, 6383, 6570/6580 and 15870 Da. The best cancer detection model was 91% accurate and included mammographic findings, age and one SELDI-identified protein: 4262 Da. The protein with the highest specificity in cancer association, 3592 Da, was identified as a β -casein-like peptide. All of the information in the model is available prior to each participant's surgery, and optimizes the identification of markers that are associated with breast cancer.

Executive summary

- Differential nipple aspirate fluid (NAF) proteomic expression exists between women/without breast cancer.
- The best sensitivity (~67%) mass (15870 Da) considering all enrolled subjects was approximately threefold more likely (odds ratio: 3.33) to be present in cancerous than noncancerous breasts. The most specific (84%) mass was 3592 Da. Mass 6570 Da was fivefold more likely (odds ratio: 0.19) to be present in noncancerous than cancerous breasts.
- Among women without pathologic nipple discharge (PND), a mass of 6383 Da had the best sensitivity (88%), and 3592 Da (87%) the best specificity.
- Since the markers associated with breast cancer differed in the PND and no-PND groups, knowing if a woman has PND is important prior to choosing which masses to search for in the NAF of a woman with a suspicious breast lesion.
- In our study, atypical or malignant cells were more readily detected in breasts containing ductal carcinoma *in situ* than invasive breast cancer. Overall, NAF cytology was highly specific but not very sensitive in determining if a breast contains cancer, and cytology only entered one of the models for cancer detection.
- The optimal model was 92% specific and 87% sensitive in determining if a breast requiring surgery had cancer.

- The best cancer detection model correctly determined if a breast contained cancer 91% of the time. The model included mammographic findings, age and the SELDI-identified protein: 4262 Da.
- The protein with the highest specificity (3592 Da) was identified as a β -casein-like peptide.
- Combining the proteomic and clinical information that are available before surgery optimizes the determination if a breast contains cancer.

Acknowledgments

Financial & competing interests disclosure

The authors received funding through NIH grant CA 95484. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Bibliography

1. Fahy BN, Bold RJ, Schneider PD, Khatri V, Goodnight JE Jr. Cost-benefit analysis of biopsy methods for suspicious mammographic lesions ; discussion 994–5. *Arch. Surg.* 136(9):990–994. [PubMed: 11529819]
2. Lehman CD, Gatsonis C, Kuhl CK, et al. MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer. *N. Engl. J. Med.* 2007; 356(13):1295–1303. [PubMed: 17392300]
3. Whelan LC, Power KA, McDowell DT, Kennedy J, Gallagher WM. Applications of SELDI-MS technology in oncology. *J. Cell. Mol. Med.* 2008; 12(5A):1535–1547. [PubMed: 18266982]
4. Sauter ER, Shan S, Hewett JE, Speckman P, Du Bois GC. Proteomic analysis of nipple aspirate fluid using SELDI-TOF-MS. *Int. J. Cancer.* 2005; 114(5):791–796. [PubMed: 15609313]
5. Sauter ER, Zhu W, Fan XJ, Wassell RP, Chervoneva I, Du Bois GC. Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. *Br. J. Cancer.* 2002; 86(9):1440–1443. [PubMed: 11986778]
6. He J, Gornbein J, Shen D, et al. Detection of breast cancer biomarkers in nipple aspirate fluid by SELDI-TOF and their identification by combined liquid chromatography-tandem mass spectrometry. *Int. J. Oncol.* 2007; 30(1):145–154. [PubMed: 17143523]
7. Noble JL, Dua RS, Coulton GR, Isacke CM, Gui GP. A comparative proteomic analysis of nipple aspiration fluid from healthy women and women with breast cancer. *Eur. J. Cancer.* 2007; 43(16):2315–2320. [PubMed: 17904354]
8. Pawletz CP, Trock B, Pennanen M, et al. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis. Markers.* 2001; 17(4):301–307. [PubMed: 11790897]
9. Pawlik TM, Fritsche H, Coombes KR, et al. Significant differences in nipple aspirate fluid protein expression between healthy women and those with breast cancer demonstrated by time-of-flight mass spectrometry. *Breast Cancer Res. Treat.* 2005; 89(2):149–157. [PubMed: 15692757]
10. Eberl MM, Fox CH, Edge SB, Carter CA, Mahoney MC. BI-RADS classification for management of abnormal mammograms. *J. Am. Board Fam. Med.* 2006; 19(2):161–164. [PubMed: 16513904]
11. American College of Radiology. BI-RADS™. 2nd Edition. VA, USA: American College of Radiology; 1995.
12. Sauter ER, Wagner-Mann C, Ehya H, Klein-Szanto A. Biologic markers of breast cancer in nipple aspirate fluid and nipple discharge are associated with clinical findings. *Cancer Detect. Prev.* 2007; 31(1):50–58. [PubMed: 17317033]

13. Agurs-Collins T, Adams-Campbell LL, Kim KS, Cullen KJ. Insulin-like growth factor-1 and breast cancer risk in postmenopausal African-American women. *Cancer Detect. Prev.* 2000; 24(3):199–206. [PubMed: 10975280]
14. Sauter ER, Ehya H, Schlatter L, MacGibbon B. Ductoscopic cytology to detect breast cancer. *Cancer J.* 2004; 10(1):33–41. discussion 15–36. [PubMed: 15000493]
15. Sauter ER, Ross E, Daly M, et al. Nipple aspirate fluid: a promising non-invasive method to identify cellular markers of breast cancer risk. *Br. J. Cancer.* 1997; 76(4):494–501. [PubMed: 9275027]
16. Li, X.; Gentlemen, R.; Lu, X., et al. SELDI-TOF Mass Spectrometry Protein Data. NY, USA: Springer; 2005.
17. Mani, DR.; Gillette, M. Proteomic Data Analysis: Pattern Recognition for Medical Diagnosis and Biomarker Discovery. NJ, USA: IEE Press; 2004.
18. Gentlemen RC, Vandal AC. Computational algorithms for censored data problems using intersection graphs. *J. Comput. Graph. Stat.* 2001; 10:403–421.
19. Yasui Y, Pepe M, Thompson ML, et al. A data-analytic strategy for protein-biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics.* 2003; 4:449–463. [PubMed: 12925511]
20. Coombes KR, Koomen JM, Baggerly KA, Morris JS, Kobayashi R. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics.* 2005; 1:41–52. [PubMed: 19305631]
21. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.* 1983; 78:316–330.
22. Meisel H. Biochemical properties of peptides encrypted in bovine milk proteins. *Curr. Med. Chem.* 2005; 12(16):1905–1919. [PubMed: 16101509]
23. Kampa M, Loukas S, Hatzoglou A, Martin P, Martin PM, Castanas E. Identification of a novel opioid peptide (Tyr-Val-Pro-Phe-Pro) derived from human α S1 casein (α S1-casomorphin, and α S1-casomorphin amide). *Biochem. J.* 1996; 319(Pt 3):903–908. [PubMed: 8920997]
24. Leiting B, Pryor KD, Wu JK, et al. Catalytic properties and inhibition of proline-specific dipeptidyl peptidases II, IV and VII. *Biochem. J.* 2003; 371(Pt 2):525–532. [PubMed: 12529175]
25. Riedl CC, Ponhold L, Flory D, et al. Magnetic resonance imaging of the breast improves detection of invasive cancer, preinvasive cancer, and premalignant lesions during surveillance of women at high risk for breast cancer. *Clin. Cancer Res.* 2007; 13(20):6144–6152. [PubMed: 17947480]

Website

101. Matrix Science. www.matrixscience.com

Figure 1.
 Venn diagram demonstrating the commonality of SELDI-TOF biomarkers in the full dataset with that of subgroups which exclude women with pathologic nipple discharge or papillomas.
 PND: Pathologic nipple discharge.

Figure 2. SELDI data from the entire set of 163 samples from IMAC chips

(A) Cancer mass spectra are shown in red, noncancer mass spectra are shown in green. The SELDI data illustrate two broad peaks averaged to identify a cancer-prevalent IMAC peak of 3592 Da. After purification of pooled cancer and noncancer samples using IMAC columns under the same conditions used for SELDI chip processing, MALDI results (B) show two well-resolved peaks (3585 and 3603 Da) present in the cancer samples. For clarity, only the region within the range of the 3592 Da SELDI peak is shown. IMAC: Immobilized metal affinity chromatography.

Figure 3. Identification of a 3602 Da SELDI peak as β -casein

(A) The QTOF MS mass spectrum of the ZIP tip elution of IMAC-purified nipple aspirate fluid. The relative intensity of peptide ions is on the y-axis and m/z ratio in amus is on the x-axis. The 5+ ion of the 3603 Da SELDI peak is 721.83 and the 4+ ion of the same peptide is 902.02. The inset shows the MS/MS (TOF product) spectrum of the 5+ ion. (B) Zoomed region of the MS/MS spectrum (TOF Product of 721.8; 5+ ion of 3603 Da peptide). *De novo* sequencing based on triply-charged fragment ions within this mass range is shown. Peaks are labeled with a m/z value and their charge states are given in brackets. amu: Atomic mass unit; m/z: Mass:charge ratio.

Table 1Patient demographics in subjects with and without cancer^{*}.

	Cancer (n = 47)	No cancer (n = 78)	p-value
Age – median (years)	61	48.5	<0.0001
Age – range (years)	39–95	22–77	
Postmenopausal [‡] (%)	37 (80)	39 (52)	0.0013
First-degree relatives with breast cancer [§] (%)	16 (34)	14 (18)	0.038
Parous (%)	43 (91.5)	69 (88.5)	0.42
Current birth control pill use (%)	0 (0)	2 (2.6)	0.53
Current hormone-replacement therapy use (%)	6 (12.8)	12 (15.4)	0.45

^{*} One individual had bilateral cancer (ductal carcinoma in situ in each breast).

[‡] Menopausal status unknown for four subjects.

[§] Family history unknown for one subject.

Table 2

Breast Imaging Reporting and Data System category and corresponding pathologic diagnosis for breasts from which nipple aspirate fluid was analyzed category.

BI-RADS category	Diagnosis	
	<i>Cancer (n)</i>	<i>Benign (n)</i>
0	1	11
1	0	16
2	1	21
3	2	6
4	27	18
5	15	1
NA *	2	42

* BI-RADS category not available because subject did not undergo mammography.
 BI-RADS: Breast Imaging Reporting and Data System; NA: Not applicable.

Table 3

Protein expression profile in nipple aspirate fluid (by size) obtained from breasts with and without cancer.

Chip	Protein peak (Da)	Positive (%) in cancer (sensitivity)	Positive (%) in normal (1 – specificity)	OR (95% CI; conditional MLE)	p-value *
		n = 48	n = 115		
IMAC30	3592	33.33	16.53	2.50 (1.08, 5.88)	0.022
	6580	41.67	66.96	0.35 (0.17, 0.75)	0.005
	6710	18.75	40.00	0.35 (0.14, 0.82)	0.011
	13005	62.50	82.61	0.35 (0.15, 0.81)	0.008
	13207	56.25	76.52	0.40 (0.18, 0.86)	0.014
		n = 35	n = 106		
CM10	6570	5.71	24.53	0.19 (0.02, 0.83)	0.015
	11747	45.71	66.04	0.44 (0.18, 1.01)	0.045
	14698	22.86	44.34	0.37 (0.13, 0.94)	0.028
	15334	42.86	21.70	(2.70 1.10, 6.67)	0.027
	15447	17.14	36.80	0.36 (0.11, 0.98)	0.037
		n = 43	n = 111		
Q10	15135	48.88	26.12	2.70 (1.23, 5.88)	0.008 [‡]
	15157	46.67	26.12	2.44 (1.12, 5.56)	0.022
	15427	17.78	39.64	0.33 (0.12, 0.81)	0.009 [‡]
	15870	66.67	36.94	3.33 (1.56, 7.69)	0.001 [‡]
	22434	15.56	46.85	0.40 (0.14, 1.03)	0.047
	22512	31.11	48.65	0.45 (0.20, 0.97)	0.033

Obtained using three different SELDI-TOF ProteinChips (IMAC30, CM10, Q10) [1,2]. Only proteins with a p-value 0.05 are listed.

* Nominal p-values based on Fisher's exact test are given.

[‡]Significant after adjustment for multiple testing within chip type using the Benjamini-Hochberg correction.

CM10: Weak cation exchange; IMAC30: Immobilized metal affinity chromatography; MLE: Maximum likelihood estimate; NAF: Nipple aspirate fluid; OR: Odds ratio; Q10: Strong anion exchange.

Table 4

Nipple aspirate fluid protein profile in breasts with or without cancer.

Chip	Protein peak (Da)	Positive (%) in cancer (sensitivity)	Positive (%) in normal (1 – specificity)	OR (95% CI)	p-value*
<i>No PND</i>		<i>n = 41</i>	<i>n = 81</i>		
IMAC30	3592	36.59	13.58	3.67 (1.35, 10.00)	0.005
	6383	87.80	98.77	0.09 (0.002, 0.87)	0.016
	6441	85.37	97.54	0.15 (0.01, 0.89)	0.017
	6580	36.59	65.44	0.31 (0.13, 0.71)	0.004 [‡]
	6621	26.83	48.15	0.40 (0.16, 0.95)	0.032
	6710	17.07	43.21	0.27 (0.09, 0.72)	0.005 [‡]
	11719	60.98	82.71	0.33 (0.13, 0.83)	0.014
	11739	21.95	43.21	0.37 (0.14, 0.93)	0.028
	11916	48.80	67.90	0.45 (0.19, 1.04)	0.050
	13005	63.41	90.12	0.19 (0.06, 0.55)	0.001 [‡]
	13207	58.54	82.71	0.30 (0.12, 0.75)	0.007
	15871	34.14	16.05	2.70 (1.02, 7.14)	0.036
<i>No PND</i>		<i>n = 29</i>	<i>n = 75</i>		
CM10	6570	3.44	29.33	0.09 (0.002, 0.60)	0.003 [‡]
	14698	27.59	49.34	0.40 (0.13, 1.08)	0.050
	15870	75.86	52.00	2.86 (1.03, 9.09)	0.029
<i>No PND</i>		<i>n = 39</i>	<i>n = 81</i>		
Q10	15135	41.03	17.30	3.33 (1.28, 8.33)	0.007 [‡]
	15157	38.46	17.28	2.94 (1.15, 7.69)	0.021
	15427	20.51	43.21	0.34 (0.12, 0.88)	0.016
	15870	64.10	28.40	4.35 (1.85, 11.11)	<0.001 [‡]
	22512	35.90	58.02	0.41 (1.85, 11.11)	0.032

Obtained using three ProteinChips: IMAC30, CM10 and Q10, excluding breasts with PND and papilloma. Only proteins with a p-value 0.05 are listed.

* Nominal p-values based on Fisher's exact test are given.

† Significant after adjustment for multiple testing within chip type using the Benjamini-Hochberg correction.

CM10: Weak cation exchange; IMAC30: Immobilized metal affinity chromatography; OR: Odds ratio; PND: Pathologic nipple discharge; Q10: Strong anion exchange.

Table 5

Ability of clinical and SELDI-TOF biomarkers to predict cancer in the breast^{*}.

Method	Data	Predictors	Cross-validation results		
			Correct (%)	Specificity (%)	Sensitivity (%)
Stepwise logistic regression	Clinical [‡]	BI-RADS, age, race, cytology	85.0	90.4	63.3
	SELDI	2016, 2753, 3252, 3485, 3592, 4135, 4263, 4279, 4471, 4786, 4800, 11236, 13005, 13373, 15124, 15157, 15391, 15870	78.9	91.2	61.3
	Both	BI-RADS, AGE, 3592, 13005, 6638, 4132, 4262	89.4	93.1	77.4
Stepwise linear discriminant	Clinical	BI-RADS, age	87.2	91.1	74.2
	SELDI	3592, 4129	80.5	99.0	19.4
	Both	BI-RADS, age, 4262	91.0	92.2	87.1
Stepwise quadratic	Clinical [‡]	BI-RADS, age, postmenopausal, PND	87.9	90.4	80.0
Discriminant	SELDI	3592, 3085	81.2	95.1	35.5
	Both	BI-RADS, Age, 13373	90.2	92.2	83.9
CART	Clinical [‡]	BI-RADS, age, postmenopausal, PND, race	79.8	90.4	46.7
	SELDI	4799, 6563, 11747, 4017, 6383	62.4	63.7	58.0
	Both	BI-RADS, age, 3592	82.7	94.1	45.2

^{*} n = 133 cases (102 noncancer, 31 cancer), unless otherwise noted.

[‡] n = 124 cases (94 noncancer, 30 cancer) due to missing data on postmenopause, PND, race or cytology.

BI-RADS: Breast Imaging Reporting and Data System; CART: Classification and regression trees; PND: Pathologic nipple discharge.

Table 6

SELDI-TOF studies evaluating nipple aspirate fluid in women with breast cancer.

Population	n	NAF collected before/after surgery	Removed abundant proteins	Predictive markers (Da)		Ref.
				Cancer	No cancer	
Healthy women and women with breast cancer	27	Unknown	no	4233, 9470	3415, 4149	[8]
Healthy women and women requiring breast surgery	33	Before and after	no	6500, 8000 [*] , 15940 [*] , 28100, 31770 [*]		[5]
Women requiring breast surgery	98	Before	no	5200, 11880, 13880, 33400		[4]
Healthy women and women with breast cancer [†]	28	Unknown	no	952, 1028, 6721, 6759	2310, 2520, 2791, 2956, 2976, 3017, 3284, 3865, 4182, 4205, 4760, 5092, 5101, 6307, 8157, 8385, 8419, 8573, 9369, 13891, 13979, 16707, 16915	[9]
Healthy women and women with breast cancer	76	Before	Yes	5277, 5284, 5304, 5323, 5330, 5363, 5477, 5944, 6036, 53668	12987, 13300, 13363, 13734	[6]
Healthy women and women with breast cancer [†]	65	Before	No		3471, 3501, 3511, 3627, 4147, 4151, 4586, 4646, 4698	[7]

^{*}These proteins were later found to be forms of hemoglobin-β chain.

[†]We were unable to determine from the manuscript if the marker was more prevalent in samples from women with breast cancer or healthy women. NAF: Nipple aspirate fluid.